Music separation guided by cover tracks: designing the joint NMF model

Nathan Souviraà-Labastie, Emmanuel Vincent, Frédéric Bimbot

MUSIC SEPARATION GUIDED BY COVER TRACKS: DESIGNING THE JOINT NMF MODEL

Nathan Souviraà-Labastie[*] *Emmanuel Vincent*[†] *Frédéric Bimbot*[‡]

[*] Université de Rennes 1, IRISA - UMR 6074, Campus de Beaulieu 35042 Rennes cedex, France
[†] Inria, Centre de Nancy - Grand Est, 54600 Villers-lès-Nancy, France
[‡] CNRS, IRISA - UMR 6074, Campus de Beaulieu 35042 Rennes cedex, France

ABSTRACT

In audio source separation, reference guided approaches are a class of methods that use reference signals to guide the separation. In prior work, we proposed a general framework to model the deformation between the sources and the references. In this paper, we investigate a specific scenario within this framework: music separation guided by the multitrack recording of a cover interpretation of the song to be processed. We report a series of experiments highlighting the relevance of joint Non-negative Matrix Factorization (NMF), dictionary transformation, and specific transformation models for different types of sources. A signal-to-distortion ratio improvement (SDRI) of almost 11 decibels (dB) is achieved, improving by 2 dB compared to previous study on the same data set. These observations contribute to validate the relevance of the theoretical general framework and can be useful in practice for designing models for other reference guided source separation problems.

Index Terms— Music separation, Joint-NMF, Cover song

1. INTRODUCTION

In signal processing, audio source separation is the task of recovering the different audio sources that compose an observed audio mixture. In the case of music, this task aims to provide signals for each instrument or voice. As original songs are rarely released in multitrack formats, this step is compulsory to open new possibilities in music post-production, *e.g.*, respatialization, upmixing and more widely in audio edition.

The complexity of the mixing process (not necessarily linear) as well as the fact that there are more sources than input channels make the demixing of professionally produced music difficult in the blind case, *i.e.*, without any prior information. Thus, blind separation shows certain limitations for professional music applications that require high audio quality [1]. Many approaches have taken several kinds of additional information into account with the objective of overcoming these limitations [2]. For instance, spatial and spectral information about the sources [3], information about the recording/mixing conditions [4], musical scores [5, 6], or even selection of spectrogram areas [7, 8, 9], potentially in an interactive way [10, 11, 12] have been proposed in the literature. It is also possible to consider *reference signals* [13] that are similar to the sources to be separated, for instance uttered by a user [14, 15, 16], or synthesized from symbolic information [15, 17] or even retrieved from a large set [18, 19].

In this paper, we focus on source separation guided by reference signals, and more precisely music separation guided by cover multitrack songs [20]. A cover song is another performance of an original song. It can differ from the original song by its musical interpretation and it is potentially performed by a different singer and with different instruments. Multitrack recordings of such covers are more likely to be found on the market than the original multitrack recording and contrary to expectations, they are (for commercial reasons) musically faithful to the original [20]. Furthermore, most separation algorithms are sensitive to initialization and using cover multitrack recordings for initialization is an efficient way to sidestep this problem [20]. All these reasons make cover guided music separation a very promising approach for high quality music separation.

In the following, rather than using the cover multitrack recordings for initialization only, we also use them to constrain the power spectrum of each source. In addition, although the considered covers are musically faithful, deformations exist between the original sources and the covers at the signal level. These deformations are significant enough not to be ignored. Here, different configurations of deformations as formalized in [13] are tested. Finally, the optimal deformation model is selected for each type of source (bass, drums, guitar, vocal ...).

The paper is organized as follows. Section 2 recalls the general model of reference guided source separation proposed in [13]. Section 3 presents a specific model adapted to the task, and its estimation procedure. Section 4 describes the data and the settings of the experiments. Section 5 reports on results illustrating the proposed contributions and provides to the reader useful advice on how to design deformation models for reference guided source separation. Section 6 draws the conclusions and gives some perspectives.

2. GENERAL FRAMEWORK

In this section, we recall the general framework for multi-channel source separation guided by multiple deformed references [13].

The observations are M multi-channel audio mixtures $\mathbf{x}^m(t)$ indexed by m and containing I^m channels. For instance, one mixture is to be separated, and the other mixtures contain the reference signals used to guide the separation process. Each mixture $\mathbf{x}^m(t)$ is assumed to be a sum of source spatial images $\mathbf{y}_j(t)$ indexed by $j \in \mathcal{J}_m$. In the Short-Time Fourier Transform (STFT) domain, this can be written as

$$\mathbf{x}^m_{fn} = \sum_{j \in \mathcal{J}_m} \mathbf{y}_{j,fn} \text{ with } \mathbf{x}^m_{fn}, \mathbf{y}_{j,fn} \in \mathbb{C}^{I^m}, \qquad (1)$$

where $f = 1, ..., F$ and $n = 1, ..., N$ are respectively the frequency and the time indexes of the STFT. We assume that the time-frequency coefficients of the source spatial images $\mathbf{y}_{j,fn}$ have a zero-

[*]Work supported by Maia Studio and Bretagne Region scholarship

mean Gaussian distribution [3] :

$$\mathbf{y}_{j,fn} \sim \mathcal{N}_{\mathbb{C}}(0, v_{j,fn}\mathbf{R}_{j,fn}) \quad (2)$$

whose covariance factors into a scalar power spectrum $v_{j,fn} \in \mathbb{R}_+$ and a spatial covariance matrix $\mathbf{R}_{j,fn} \in \mathbb{C}^{I^m \times I^m}$. The spatial covariance matrices model the spatial characteristics of the sources, like phase or intensity difference between channels. The power spectrogram of each source j is denoted as $V_j = [v_{j,fn}]_{fn} \in \mathbb{R}_+^{F \times N}$. Each V_j is split into the product of an excitation spectrogram V_j^e and a filter spectrogram V_j^ϕ. The excitation spectrogram (resp. the filter spectrogram) is decomposed by a Non-negative Matrix Factorization (NMF) into a matrix of spectral patterns $W_j^e \in \mathbb{R}_+^{F \times D^e}$ (resp. $W_j^\phi \in \mathbb{R}_+^{F \times D^\phi}$) and a matrix of temporal activations $H_j^e \in \mathbb{R}_+^{D^e \times N}$ (resp. $H_j^\phi \in \mathbb{R}_+^{D^\phi \times N}$). D^e (resp. D^ϕ) denotes the number of spectral patterns used in the NMF decomposition of the excitation (resp. filter) part. This results in the following decomposition :

$$V_j = V_j^e \odot V_j^\phi = W_j^e H_j^e \odot W_j^\phi H_j^\phi \quad (3)$$

where $\odot$ denotes point wise multiplication.

As the different audio mixtures are composed of similar sources, the matrices W and H can be *shared* (*i.e.*, jointly estimated) between a given source $j \in \mathcal{J}^m$ and one or more related sources $j' \in \mathcal{J}^{m'}$ with $m' \neq m$. A reference signal is by nature deformed : if there were no deformation, the reference would be equal to the true source. These differences can be taken into account by partly sharing the model parameters (*e.g.*, sharing only H^e) and/or by using transformation matrices $T_{jj'}$ (between sources j and j'). For instance, the sharing of excitation properties between two sources j and j' is modeled by one of the three following configurations, depending on the sharing of W^e and H^e :

$$V_{j'}^e = T_{jj'}^{fe} W_j^e H_{j'}^e \quad (4)$$

$$V_{j'}^e = W_{j'}^e H_j^e T_{jj'}^{te} \quad (5)$$

$$V_{j'}^e = T_{jj'}^{fe} W_j^e T_{jj'}^{de} H_j^e T_{jj'}^{te} \quad (6)$$

As it would be redundant, $T_{jj'}^{de}$ only appears if the corresponding W_j^e, H_j^e are *shared*.

3. MODEL AND ALGORITHM

In this section, we introduce the specific framework used to model the cover guided music separation problem in reference to the general framework recalled in Section 2 from which we removed the excitation-filter decomposition and the multi-channel formulation. Here, the reference mixtures are the different cover tracks and contain a single source. This configuration with isolated references leads to a very efficient initialization (detailed in Section 3.2)

3.1. Proposed model

Here, we consider that $\mathbf{x}^1(t)$ is the song to be separated, and $\mathbf{x}^m(t)$ for $m > 1$ are the different tracks of the cover version used to guide the separation process. All mixtures are single channel, and the mixture $\mathbf{x}^m(t)$ for $m > 1$ are assumed to contain only one source.

Each V_j is decomposed by a NMF into a matrix of spectral patterns $W_j \in \mathbb{R}_+^{F \times D}$ and a matrix of temporal activations $H_j \in \mathbb{R}_+^{D \times N}$.

$$V_j = W_j H_j \quad (7)$$

D denotes the number of spectral patterns used in the NMF decomposition. Hereafter, we only consider frequency and dictionary transformation matrices that are now denoted $T_{jj'}^f \in \mathbb{R}_+^{F \times F}$, and $T_{jj'}^d \in \mathbb{R}_+^{D \times D}$. As each track of the two versions are sufficiently aligned in time, we do not consider T^t matrices that induce unwanted smoothing effects. Thus, the related sources are modeled using equation (6):

$$V_{j'} = T_{jj'}^f W_j T_{jj'}^d H_j \quad (8)$$

It can be noticed that this formulation leaves the possibility to set these transformation matrices either in the reference model ($j \in \mathcal{J}^1$ and $j' \in \mathcal{J}^{m'}$, $m' \neq 1$) or in the source model ($j' \in \mathcal{J}^1$ and $j \in \mathcal{J}^{m'}$, $m' \neq 1$). See Tables 3 and 4 in Section 5 for concrete cases. For both T^f and T^d matrices, we consider two possible initializations:

- `Diag` : Identity matrix
- `Full` : The sum of an identity matrix and a random matrix drawn from a rectified Gaussian distribution

3.2. Parameter estimation

Here, we present a method for parameter estimation in the maximum likelihood (ML) sense. In the single-channel case, maximizing the log-likelihood is equivalent to minimizing the Itakura-Saito divergence [21]:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{m=1}^{M} \sum_{f,n=1}^{F,N} d_{IS}(X_{fn}^m | V_{fn}^m) \quad (9)$$

where θ is the set of parameters to be estimated, *i.e.*, the matrices W, H and T. $X^m = [|\mathbf{x}_{fn}^m|^2]_{fn}$ and $V^m = \sum_{j \in \mathcal{J}^m} V_j$ are respectively the observed and estimated power spectrograms, and $d_{IS}(a|b) = a/b - \log(a/b) - 1$ is the Itakura-Saito divergence. A common way to estimate the parameters is the use of a multiplicative gradient descent approach [21] in which each parameter is updated at each iteration without increasing criterion (9) [22]. The update of one parameter consists in multiplying it by the ratio of the negative and positive parts of the derivative of the criterion with respect to this parameter. Different multiplicative updates (MU) are derived for each parameter. Examples of such updates are given in [18].

The results of MU depend on initialization. With respect to blind source separation, reference guided separation provides better initial parameter values taking advantage of the provided references. For instance, in our case, we can use MU to minimize the following criterion :

$$\hat{\theta}_{\text{ref}} = \underset{\theta_{\text{ref}}}{\operatorname{argmin}} \sum_{m=2}^{M} \sum_{f,n=1}^{F,N} d_{IS}(X_{fn}^m | \tilde{V}_{fn}^m) \quad (10)$$

where $\tilde{V}_{fn}^m = W_j H_j$ with $j \in \mathcal{J}^m$ ($m > 1$), and θ_{ref} is the set of corresponding W_j and H_j parameters. This is especially efficient when there is a single source in the reference signal, as here. In the experiments, we will distinguish the following successive initialization and algorithmic stages :

- *Pre-NMF* : MU that try to minimize the criterion (10),
- *Joint-NMF* : MU that try to minimize the criterion (9).

In all experiments, 50 iterations of *Pre-NMF* are beforehand performed and the final source estimates are obtained using an adaptive Wiener filter.

Title	Track names
I Will Survive	Bass, Brass, Drums, Guitar, Strings, Vocal.
Pride and Joy	Bass, Drums, Guitar, Vocal.
Rocket Man	Bass, Choirs, Drums, Others, Piano, Vocal.
Walk this Way	Bass, Drums, Guitar, Vocal.

Table 1. Cover multitrack data set

	Joint-NMF		**SDRI**
	Source	Reference	**mean**
Result in [20]	WH		8.98
Reproduction of [20]	WH		8.74
Only *Pre-NMF*			10.06
Joint-NMF	WH	WH	10.27

Table 2. SDRI (dB) compared with a previous study [20].

Init		*Joint-NMF*		**SDRI**
T^f	T^d	Source	Reference	**mean**
		WH	WH	10.27
`Full`		WH	T^fWH	10.08
`Full`		T^fWH	WH	9.23
`Diag`		T^fWH	WH	10.09
`Diag`		WH	T^fWH	**10.35**
	`Diag`	WT^dH	WH	9.25
	`Diag`	WH	WT^dH	9.88
	`Full`	WH	WT^dH	9.79
	`Full`	WT^dH	WH	**10.64**

Table 3. SDRI (dB) for different configurations.

4. EXPERIMENTS

4.1. Data

In order to compare our results with previous study, we use the same data set and equivalent settings as [20]. Both original and cover multitracks are available in order to evaluate the separation. The number of cover tracks is the same as the number of sources to be separated, as each track is used as reference for the related source. They are also used in the mirror configuration, *i.e.*, the cover is to be separated and the original multitracks are used as references. Experiments are conducted on 30 second examples typically consisting of half of a verse and half of a chorus. The considered tracks of four songs are listed in Table 1. Some examples are available online[1].

4.2. Settings

We here make an exhaustive list of settings that vary from [20] and refer the reader to [20] for other common details. To simplify the analysis of the results, the two channels are summed into a single one. We used 50 iterations for *Pre-NMF* and *Joint-NMF* steps instead of 500. The number of components D is kept to 50.

5. RESULTS

The quality of the estimated sources is evaluated in terms of signal-to-distortion ratio improvement (SDRI) that is the difference between the output SDR [23] and the input SDR. The input SDR is defined as the power ratio between the source to be estimated and the mixture to be separated. It is given for each type of source in Table 4. The samples that we selected lead to an input SDR mean of -8.44 dB instead of -7.60 dB in [20].

As we work with MU, we underline that zeros in the parameters remain unchanged over the iterations. Hence, a matrix initialized at identity will stay diagonal over the iterations. Moreover, T matrices are not present during the *Pre-NMF*.

5.1. Comparison with a previous study

In [20], the cover multitrack signals are only used to initialize the source parameters W and H. Here, we decided to share these parameters between the source and the reference hence the reference signals are used during the global estimation stage too. The results are shown in Table 2.

First, we reproduce the experiments in [20] with the differences of settings presented in Section 4.2. An equivalent SDRI mean is obtained compared to [20] for the case where the parameters are not shared (8.74 dB instead of 8.98 dB). The previous configuration leads in fact to an important decrease of the SDRI mean, compared to what is obtained if the sources are reconstructed directly after the *Pre-NMF* (10.06 dB). This can be explained by the great level of similarity between the covers and the original tracks. Conversely, sharing the NMF parameters during the final estimation (*Joint-NMF*) guarantees not to move away too much from this relevant starting point while getting closer to a solution that fits better the original tracks. In our case, a marginal improvement is observed (10.27 dB).

These first results show the strong similarity between each original track and its related cover track. In that case, sharing W and H during the *Joint-NMF* estimation is the most relevant method even without considering any deformations.

5.2. Designing the deformation model

Here, we analyze whether the transformation matrices are useful in the reference or the source model. The comparison of different initializations of frequency and dictionary deformations matrices is done as well. Results of exhaustive experiments are displayed in Table 3.

Bold values indicate improvements compared to a fully shared *Joint-NMF* (10.27 dB). It can be noticed that in those two cases the number of non zero coefficients Z in T is of the same order of magnitude ($Z = D^2 = 2500$ for a `Full` T^d and $Z = F = 1025$ for a `Diag` T^f), while Z vary from $D = 50$ to $F^2 \approx 10^6$.

We also observe that for almost all cases, SDRIs are always higher when T matrices are positioned in the reference model. This can be explained by the fact that the final signals are reconstructed based on the source model, and T matrices can induce abrupt changes. Conversely, inserting a `Full` T matrix in the reference model instead of in the source model would distort more the output of the *Pre-NMF*, as the product of W_j and H_j estimated during the *Pre-NMF* step try to fit the reference signal. So, it is difficult to distinguish which effect is predominant, especially since the number of non-zero coefficients has also an impact.

One can remember that when inserting transformation matrices it is important not to distort the output of *Pre-NMF*. For instance in Table 3, it is the case for values in bold.

[1] http://speech-demos.gforge.inria.fr/source_separation/icassp2015/

	Number of tracks		4	4	3	4	1	1	1	1	1
	Input SDR (dB)		-7.42	-7.17	-9.98	-4.18	-12.34	-9.75	-12.48	-18.64	-10.55
Joint-NMF		**SDRI**	**Bass**	**Drums**	**Guitar**	**Vocal**	**Choirs**	**Others**	**Piano**	**Brass**	**Strings**
Source	Reference	**mean**									
		10.06	9.33	9.02	9.71	9.60	13.70	9.79	10.80	16.25	9.80
WH	WH	10.27	9.26	9.28	9.82	10.24	13.23	10.27	11.11	15.67	10.62
WH	T^fWH	10.35	9.39	9.29	9.69	10.22	**13.85**	***9.98***	***12.08***	15.68	**10.71**
WT^dH	WH	10.64	9.23	9.94	**10.80**	**10.73**	13.01	10.07	11.36	15.80	10.61
WT^dH	T^fWH	10.66	**9.88**	**10.48**	9.41	10.44	12.74	10.01	12.24	**16.66**	10.11
Best		10.92	9.66	10.58	10.35	10.85	13.73	10.74	11.91	14.91	11.71

Table 4. SDRI (dB) for each type of source.

5.3. Specific source model

In this last set of experiments, SDRI for each type of source are given in Table 4 for different configurations. The combination of T^d and T^f gives interesting results, especially for the drums and the bass. Moreover, we observe for each type of source clear differences between models while the SDRI means are similar. In a final experiment, referred to as "**Best**", the most proper configurations are chosen for each type of source (indicated in bold in Table 4). Note that values in italic and bold were chosen based on the experiment with T^f (`Diag`) in the source model that shows promising results for these two sources. The result (10.92 dB) shows the relevance of defining specific models depending on the source type.

It should also be noted that using transformation matrices for only one source of a song leads to a huge decrease of separation quality for that source. This is certainly due to the estimation algorithm that we used. So, defining specific source models is indeed interesting but the number of parameters should be balanced between the different models.

6. CONCLUSION

In this paper, we have addressed the problem of cover tracks guided music separation with a previously defined general framework for *audio source separation using multiple deformed references*. This study leads us to consider specific aspects of the general framework, in particular dictionary deformations. Several deformation models were tested for different types of source (Bass, Drums, Guitar, Vocal ...) which lead to 2 dB improvement compared to a previous study on the same data set. The results obtained in this article may be useful in other scenarios of reference guided audio source separation.

As this paper entirely focused on spectral modeling, considering spatial information would be an asset, for instance using [13]. However, it will be needed to compensate the potential differences of spatialization between each source and its related cover track.

7. REFERENCES

[1] Emmanuel Vincent, Nancy Bertin, Rémi Gribonval, and Frédéric Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, May 2014.

[2] Antoine Liutkus, Jean-Louis Durrieu, Laurent Daudet, and Gaël Richard, "An overview of informed audio source separation," in *Proc. International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS)*, Paris, France, 2013, pp. 1–4.

[3] Alexey Ozerov, Emmanuel Vincent, and Frédéric Bimbot, "A general flexible framework for the handling of prior information in audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1118 – 1133, 2012.

[4] Ngoc Q. K. Duong, Emmanuel Vincent, and Rémi Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

[5] Umut Simsekli, Y. Kenan Yilmaz, and A. Taylan Cemgil, "Score guided audio restoration via generalised coupled tensor factorisation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 5369–5372.

[6] Sebastian Ewert, Bryan Pardo, Meinard Müller, and Mark D. Plumbley, "Score-informed source separation for musical audio recordings: An overview," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 116–124, May 2014.

[7] Alexey Ozerov, Cédric Févotte, Raphael Blouet, and Jean-Louis Durrieu, "Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, 2011, pp. 257–260.

[8] Jean-Louis Durrieu and Jean-Philippe Thiran, "Musical audio source separation based on user-selected F0 track," in *Proc. 10th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel-Aviv, Israel, 2012, pp. 438–445.

[9] Derry FitzGerald, "User assisted separation using tensor factorisations," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug 2012, pp. 2412–2416.

[10] Benoit Fuentes, Roland Badeau, and Gaël Richard, "Blind harmonic adaptive decomposition applied to supervised source separation," in *Proc. 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, Aug 2012, pp. 2654–2658.

[11] Nicholas J. Bryan and Gautham J. Mysore, "Interactive refinement of supervised and semi-supervised sound source separation estimates.," in *Proc. IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, 2013, pp. 883–887.

[12] Ngoc Q. K. Duong, Alexey Ozerov, Louis Chevallier, and Joël Sirot, "An interactive audio source separation framework based on non-negative matrix factorization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italie, May 2014, pp. 1567–1571.

[13] Nathan Souviraà-Labastie, Anaik Olivero, Emmanuel Vincent, and Frédéric Bimbot, "Multi-channel audio source separation using multiple deformed references," Inria research report, Oct. 2014.

[14] Paris Smaragdis and Gautham J. Mysore, "Separation by humming : User-guided sound extraction from monophonic mixtures," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, 2009, pp. 69 – 72.

[15] Luc Le Magoarou, Alexey Ozerov, and Ngoc Q. K. Duong, "Text-informed audio source separation using nonnegative matrix partial co-factorization," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2013)*, Southampton, United Kingdom, Sept. 2013.

[16] Romain Hennequin, Juan José Burred, Simon Maller, and Pierre Leveau, "Speech-guided source separation using a pitch-adaptive guide signal model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italie, May 2014, pp. 6672–6676.

[17] Joachim Fritsch and Mark D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 888–891.

[18] Nathan Souviraà-Labastie, Anaik Olivero, Emmanuel Vincent, and Frédéric Bimbot, "Audio source separation using multiple deformed references," in *Proc. 22nd European Signal Processing Conference (EUSIPCO)*, Lisboa, Portugal, September 2014.

[19] Dalia El Badawy, Ngoc Q. K. Duong, and Alexey Ozerov, "On-the-fly audio source separation," in *the 24th IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2014)*, Reims, France, Sept. 2014.

[20] Timothée Gerber, Martin Dutasta, Laurent Girin, and Cédric Févotte, "Professionally-produced music separation guided by covers," in *Proc. International Conference on Music Information Retrieval (ISMIR)*, Porto, Portugal, 2012, pp. 85–90.

[21] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[22] Cédric Févotte and Jérôme Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[23] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.